Statistics Canada | Statistique Canada

Methodology Branch

Census & Household Survey
Methods Division

Direction de la méthodologie

Division des méthodes de recensement
et d'enquêtes ménages

Canada

S-179F
C.2

LA STRATIFICATION DANS L'ENQUETE SUR LA POPULATION ACTIVE DU CANADA

J.D. Drew, Y. Bélanger, et P. Foy[1]

Statistique Canada

DMREM 85-070F

La stratification dans l'enquête sur la population active du Canada

J.D. DREW, Y. BÉLANGER, et P. FOY[1]

RÉSUMÉ

L'utilisation d'un algorithme de classification multi-variée pour effectuer la stratification pour l'enquête sur la population active est décrite. L'algorithme élaboré par Friedman et Rubin (1967), est modifié de manière à traiter la formation de strates géographiquement contiguës, et à faire la délimitation d'unités primaires d'échantillonnage (U.P.É.) hétéroqènes mais compactes à l'intérieur des strates. Des études portant sur les variables de stratification, la robustesse de la stratification au fil du temps, et le type de stratification sont décrites.

MOTS CLÉS: Algorithme de classification multi-variée; stratification géographique; enquête permanente.

1. INTRODUCTION

L'enquête sur la population active du Canada (E.P.A.) est remaniée après chaque recensement décennal de la population et des logements. Dans le cadre du remaniement suivant le recensement de 1981, on a mené un programme intensif de recherche sur divers aspects du plan de sondage (Singh, Drew, et Choudhry; 1984). Ce rapport décrit la partie du programme de recherche portant sur les méthodes de stratification.

Étant donné que l'E.P.A. est utilisée, non seulement pour fournir de l'information sur les caractéristiques de la population active, mais aussi comme plan de sondage général pour diverses autres enquêtes-ménages, l'un des principaux objectifs du remaniement était d'accroître la flexibilité de l'E.P.A. pour des applications générales. La stratification a été considérée comme un moyen d'améliorer l'efficacité du plan d'échantillonnage en ce qui a trait aux

[1] J.D. Drew et Y. Bélanger, Division des méthodes de recensement et d'enquêtes-ménages, et P. Foy, Division des méthodes d'enquête pour les entreprises, Statistique Canada, Ottawa (Ontario), Canada K1A 0T6.

applications générales, de même que pour des variables d'intérêt particulier pour L'E.P.A., en adoptant des méthodes plus rigoureuses que dans l'ancien plan.

On a donc décidé de considérer l'utilisation d'algorithmes de classification multi-variée pour stratifier et mettre en grappes et de les comparer avec les méthodes utilisées dans l'ancien plan de sondage. On a choisi un algorithme non-hiérarchique élaboré par Friedman et Rubin (1967), en nous fiant aux résultats d'évaluations des différents algorithmes faites par Judkins et Singh (1981) dans le cadre de remaniement du Current Population Survey du U.S. Bureau of Census. La description de l'algorithme de base, et des extensions que nous avons développées se retrouve dans la section 2.

Les sections 3 et 4 décrivent les études d'évaluation et la stratification finalement adoptée dans les deux grands types de secteurs du plan de sondage de l'E.P.A., soient les Unités Non Auto Représentatives (U.N.A.R.) et les Unités Auto Représentatives (U.A.R.). La section 4 décrit également comment on a adapté l'algorithme pour délimiter les unités primaires d'échantillonnage à l'intérieur des strates U.N.A.R.

On conclut dans la section 5 avec quelques observations sur la possibilité d'adapter le système développé à d'autres applications.

2. ALGORITHME DE STRATIFICATION

L'algorithme de base utilisé pour la stratification est un algorithme multi-varié, non-hiérarchique élaboré par Friedman et Rubin (1967). Ce choix repose sur les résultats d'études faites par Judkins et Singh (1981) et Kostanich, Judkins, Singh et Schantz (1981), qui ont évalué plusieurs algorithmes de stratification pour le Current Population Survey du U.S. Bureau of the Census.

Ces derniers ont modifié la fonction objective de l'algorithme dans le cas de l'échantillonnage avec probabilité proportionnelle à la taille (PPT), et nous avons ajouté la capacité de formuler des strates compactes et contiguës. Une

description plus complète de ce qui suit se retrouve dans Foy (1984).

2.1 La fonction objective de l'algorithme

L'algorithme vise à partitionner les unités de stratification (secteurs de dénombrement du recensement) dans des strates les plus homogènes possible à l'égard de plusieurs variables d'intérêt, c'est-à-dire en minimisant les sommes des carrés à l'intérieur de chaque strate.

Les expressions pour les sommes des carrés dans le cas de l'échantillonnage avec PPT suivent après l'introduction de la notation utilisée:

L = nombre de strates à former

N = nombre total d'unités (secteurs de dénombrement)

N_k = nombre d'unités dans le groupe (strate) k; $(N_1 + N_2 + \ldots + N_L = N)$,

T_{jk} = mesure de la taille de l'unité j du groupe k,

$T_{.k}$ = mesure de la taille du groupe k,

$T..$ = taille totale,

${}_iX_{jk}$ = valeur observée de la variable i pour l'unité j du groupe k,

${}_iX_{.k}$ = total des valeurs observées de la variable i dans le groupe k,

${}_iX..$ = total des valeurs observées de la variable i,

W_i = facteur de pondération de la variable i (voir section 2.4 pour plus de détails).

p = nombre de variables d'intérêt

Ainsi, l'expression de la somme des carrés totale, avec PPT, de la variable i est donnée par

$$SCT_i = \sum_{k=1}^{L} \sum_{j=1}^{N_k} \frac{T_{jk}}{T..} \left(\frac{T..}{T_{jk}} {}_iX_{jk} - {}_iX.. \right)^2.$$

Ceci est aussi l'expression de la variance de l'estimateur de ${}_iX..$ lorsqu'une unité est choisie avec PPT. La somme des carrés totale pondérée sur toutes les variables est donc

$$SCT = \sum_{i=1}^{p} W_i \, SCT_i .$$

Quant aux sommes des carrés intra-groupes et inter-groupes, elles sont obtenues respectivement par les expressions suivantes:

$$SCW_i = \sum_{k=1}^{L} \frac{T..}{T_{.k}} \sum_{j=1}^{N_k} \frac{T_{jk}}{T_{.k}} \left(\frac{T_{.k}}{T_{jk}} \, {}_iX_{jk} - {}_iX_{.k}\right)^2$$

et

$$SCB_i = \sum_{k=1}^{L} \frac{T_{.k}}{T..} \left(\frac{T..}{T_{.k}} \, {}_iX_{.k} - {}_iX..\right)^2 .$$

Leurs sommes des carrés pondérées sur toutes les variables sont données respectivement par

$$SCW = \sum_{i=1}^{p} W_i \, SCW_i$$

et

$$SCB = \sum_{i=1}^{p} W_i \, SCB_i .$$

La somme des carrés intra-groupes de la variable i, SCW_i, est aussi l'expression de la variance de l'estimateur de ${}_iX..$ lorsqu'une strate est choisie avec PPT et subséquemment une unité de cette strate est choisie avec PPT.

Comme d'habitude, nous avons le résultat suivant:

$$SCT_i = SCW_i + SCB_i, \quad (i=1,\ldots, p)$$

et

$$SCT = SCW + SCB.$$

La fonction objective du programme de stratification se trouve à être SCW, la somme des carrés intra-groupes pondérée sur toutes les variables. Elle doit être minimisée. On définit l'indice de stratification, I, comme étant:

$$I_i = 100 \quad \frac{SCB_i}{SCT_i} \qquad i = 1, \ldots, p$$

Une valeur élevée de l'indice indique une bonne stratification.

2.2 Recherche de la meilleure classification

Afin d'identifier la meilleure classification, une méthode consisterait à générer toutes les partitions possibles de "N" unités en "L" groupes. Il suffirait ensuite de retenir celle qui minimise la fonction objective. Cette tactique est rarement faisable puisque le nombre de partitions possibles peut être exagérément grand.

Friedman et Rubin (1967) suggèrent l'algorithme suivant. Commençons d'abord avec une partition quelconque des "N" unités en "L" groupes. Considérons maintenant le transfert d'une unité à un groupe autre que celui auquel elle appartient. Cette unité sera transférée au groupe qui apportera la plus forte réduction de la fonction objective. Toutefois, l'unité demeurera dans son groupe original si aucun tranfert n'apporte de réduction. En se servant maintenant de la partition ainsi engendrée, traitons la deuxième unité de la même manière, ensuite la troisième, jusqu'à la $N^{ième}$. L'application de cette procédure à chaque unité devient une itération que les auteurs appellent "hill-climbing pass". Après plusieurs itérations, l'algorithme atteint un point où aucun transfert de quelqu'unité que ce soit ne permet de réduire la fonction objective. Ce point est dit un minimum local de la fonction objective parce qu'il dépend de la partition initiale utilisée. Une autre partition initiale aurait pu atteindre un autre point où la valeur de la fonction objective est encore plus faible. Pour aller au-delà du minimum local, Friedman et Rubin décrivent deux procédures, soient le "forcing pass" et le "reassignment pass". En appliquant leur algorithme aux données décrites dans leur article, ils obtiennent la plus haute valeur connue de la fonction objective 10 fois sur 14 tentatives en utilisant différentes partitions initiales (ils utilisent une

autre fonction objective qui est maximisée). Avec des données moins bien structurées, la plus haute valeur est atteinte 3 fois sur 11, sans toutefois l'assurance d'avoir obtenu la solution optimale. À leur avis, les méthodes du "forcing pass" et du "reassignment pass" ne sont utiles qu'à l'occasion. Ils ont plus confiance aux résultats obtenus en utilisant plusieurs partitions initiales. Ce point de vue est soutenu par Judkins et Singh (1981). On a donc décidé d'utiliser la technique de plusieurs partitions initiales.

Parce que l'algorithme ne déplace qu'une unité à la fois, le calcul de la fonction objective en est simplifié. Après le calcul initial de la fonction objective, il suffit de recalculer la contribution à la fonction objective des deux groupes impliqués dans le transfert de l'unité en jeu.

2.3 La contiguïté

Dans les plans de sondage antérieurs de l'E.P.A., on avait adopté des strates formées d'unités géographiques contiguës, c'est-à-dire que chaque unité d'une strate donnée devait toucher à au moins une autre unité de la même strate. Une des raisons principales était une présomption que de telles strates préserveraient l'efficacité du plan d'échantillonnage sur une plus longue période de temps que dans le cas des strates formées d'unités non-contiguës.

En vue d'évaluer cette présomption et d'adopter la meilleure stratification possible, nous avons considéré deux moyens de tenir compte de la géographie dans la stratification. La première méthode est élaborée par Dahmström et Hagnell (1978), et consiste en l'utilisation de centroïdes comme variables d'intérêt. Cette méthode fait appel à deux variables géographiques (centroïdes) qui sont des transformations de longitude et de latitude. Cette méthode donne des strates compactes, c'est-à-dire des strates où la distance entre les unités est rendue minimale par la minimisation de la somme des carrés intra-groupe usuelle des centroïdes. Cette minimisation est tempérée par la minimisation des autres variables d'intérêt. Aussi il n'y a pas de garantie que les strates soient ainsi formées d'unités contiguës.

L'autre méthode, que nous appelons l'approche de vecteurs de contiguïté, est

nouvelle. Elle garantie des strates contiguës, mais pas nécessairement compactes. Des études décrites dans la section 3, ont porté sur l'utilisation de l'une ou l'autre des méthodes seulement ou d'une combinaison des deux méthodes.

2.3.1 Vecteurs de contiguïté

Afin d'assurer la formation de strates contiguës, on a procédé de la façon suivante. On fait l'optimisation comme décrite dans la section précédente mais en commençant dans ce cas avec une partition initiale qui est contiguë, et en permettant le déplacement de l'unité "j" de la strate A à la strate B, seulement si en plus de réduire les sommes des carrés, les conditions suivantes sont respectées:

(i) l'unité "j" est contiguë à une unité de la strate B

(ii) le déplacement de l'unité "j" à la strate B ne dérangera pas la contiguïté de la strate A.

Pour vérifier ces deux conditions, il est essentiel de connaître les liens de contiguïté entre les unités. Par conséquent, à chaque unité doit être assigné un vecteur de contiguïté qui contient la liste des unités qui y sont contiguës.

La première condition est facile à vérifier. Pour s'assurer que l'unité "j" est contiguë à une unité de la strate B, il suffit de trouver dans son vecteur de contiguïté une unité qui appartient à la strate B.

La deuxième condition est plus difficile à vérifier. Le principe est qu'une strate est dite contiguë si chaque paire d'unités de cette strate peut être reliée par une chaîne continue d'unités provenant de cette même strate. Soit l'unité "j" que l'on veut déplacer de la strate A à la strate B. Il est donc nécessaire de chercher, pour chaque paire d'unités du vecteur de contiguïté de l'unité "j" appartenant à la strate A, un autre lien parmi les unités de la strate A. A ce stage, le problème se compare à celui de trouver son chemin

dans un labyrinthe.

Un algorithme a aussi été conçu pour créer de façon aléatoire des partitions initiales contiguës.

2.4 Pondération des variables

Les facteurs de pondération revêtent une importance toute particulière. Ce sont eux qui établissent l'apport de chaque variable à la classification.

Il est habituellement préférable de standardiser les variables, en rendant les facteurs de pondération inversement proportionnels à la somme des carrés totale de chaque variable. Cette standardisation permet d'obtenir un apport comparable de chaque variable à la classification.

Si après la standardisation on veut qu'une ou plusieurs variables se voient accorder plus d'importance relativement aux autres variables dans l'optimisation, on peut le faire en spécifiant un poids supérieur à 1 (la normale). Par exemple, une variable avec un poids de 2 aurait une importance double. Comme décrit dans la section 3.2 on a dû essayer plusieurs combinaisons de poids pour les variables à caractère géographique et non-géographique dans le but d'obtenir des strates compactes sans trop affecter la minimisation des autres variables.

3. LA STRATIFICATION DANS LES UNITÉS NON AUTOREPRÉSENTATIVES

3.1 Ancienne conception (Platek et Singh 1976)

Pour l'E.P.A., chacune des dix provinces canadiennes est divisée en un certain nombre de régions économiques (R.É.), composées de secteurs ayant des structures économiques semblables. Les frontières des R.É. sont déterminées en consultation avec les provinces. Ces R.É. sont considérées comme des strates primaires. L'étape de stratification suivante est la partition de chaque R.É. en unités autoreprésentatives (U.A.R.) et unités non autoreprésentatives

(U.N.A.R.). Les unités autoreprésentatives sont les villes où l'échantillon prévu est assez important pour constituer au moins une tâche d'interviewer; la partie N.A.R. comprend le reste de la R.É. Des plans d'échantillonnage différents sont suivis dans les U.A.R. et U.N.A.R., parce que la population dans les U.N.A.R. est beaucoup plus dispersée, rendant nécessaire un plus grand nombre d'étapes d'échantillonnage. Pour ces mêmes raisons, on retient le concept des U.A.R. et U.N.A.R. dans le remaniement.

Dans l'ancien plan, on a stratifié la partie N.A.R. de chaque R.É. en un maximum de 5 strates contiguës ayant une population entre 36,000 et 75,000 personnes, par rapport aux caractéristiques principales de la population du recensement de 1971, comme décrit ci-dessous, et comme élaboré en plus de détail par Platek et Singh (1976).

La population active a été répartie en 7 catégories selon l'industrie. Dans chaque R.É., on a choisi les 3 catégories les plus importantes selon certains critères spécifiques. Les municipalités groupées, qui représentent les régions géographiques comprises dans une municipalité rurale, et qui, par conséquent, renferment souvent des municipalités urbaines géographiquement plus petites, ont été les unités de stratification. En comparant, pour chacune de ces unités, les proportions de la population active appartenant à chacune des trois catégories, avec les mêmes proportions calculées au niveau de la R.É., on a pu identifier les unités montrant une certaine ressemblance entre elles pour les regrouper dans des strates. Cette comparaison a été effectuée visuellement à l'aide de graphiques. En général, il a fallu procéder à certains rajustements pour satisfaire les exigences relatives à la taille et à la contiguïté des strates.

A l'intérieur de chaque strate, de 12 à 15 U.P.É. ont été formées, de façon à être semblables à la strate relativement aux variables de stratification, et relativement au rapport entre la population rurale et urbaine. Les parties rurales des U.P.É. ont été formées de S.D. contigus, et les parties urbaines ont été choisies de façon à être aussi près géographiquement que possible de la partie rurale. Les tailles des strates et des U.P.É. ont été déterminées

de sorte qu'avec deux U.P.É. sélectionnées par strate, la taille de l'échantillon prévue soit équivalente à une tâche d'interviewer. Suivant ces critères, dépendant de la province, la population des U.P.É. variait entre 3,000 et 5,000 personnes. A l'intérieur des U.P.É., l'échantillonnage a été fait en 2 ou 3 étapes.

3.2 Les études sur la stratification au moment du remaniement

Le but de nos études était de tirer des conclusions permettant de prendre des décisions relatives aux aspects suivants de la stratification: variables à utiliser, type de strates (entièrement rurales, entièrement urbaines ou mixtes), et importance à accorder à la contiguïté. Étant donné le temps très limité pour les études avant le moment de la formation des nouvelles strates et U.P.É., et l'attente générale que des strates formées d'unités contiguës seraient préférables au fil du temps à des strates formées d'unités non-contiguës, les deux premiers aspects ont été jugés prioritaires.

En ce qui concerne la contiguïté, on a dû expérimenter pour trouver le meilleur moyen de l'atteindre, soit par les vecteurs de contiguïté, les centroïdes ou une combinaison des deux. Cependant, après le remaniement, on a entrepris une étude plus approfondie examinant le choix entre des strates contiguës et des strates non-contiguës.

3.2.1 Étude sur les variables et le type de stratification

Une contrainte de la méthode de stratification utilisée dans l'ancien plan de sondage était le nombre limité de variables de stratification que l'on pouvait considérer (3 par R.É.).

Avec le nouvel algorithme, on s'est libéré de cette contrainte. En plus des sept variables d'industrie, on a voulu voir l'effet causé par l'utilisation de variables reliées au sujet de l'enquête, telles que: taux de personnes occupées, taux de chômage et niveau de revenu, ainsi que par des caractéristiques telles que: niveau d'éducation, type de logement et nombre total d'individus. Ces dernières caractéristiques se sont révélées très efficaces dans des études

semblables menées par le U.S. Bureau of the Census pour le Current Population Survey.

Le tableau 1 décrit les diverses options étudiées en ce qui a trait au choix de variables.

En ce qui concerne le type de stratification, on a décider d'étudier l'effet de former des strates séparément pour les parties rurales et urbaines des R.É., comme une alternative à la méthode mixte de l'ancien plan.

Les contraintes du plan de sondage selon lesquelles les U.P.É. doivent avoir des populations presque égales, tandis que le rapport entre la population rurale et la population urbaine doit demeurer presque le même pour chaque U.P.É., ont résulté par le passé en un manque de contiguïté entre les parties rurales et urbaines des U.P.É. Ceci a mené à une érosion de la présumée correspondance entre l'U.P.É. et la tâche d'intervieweur. On a pensé que les strates séparées en parties rurales et urbaines, qui pourraient être sous-stratifiées d'une façon optimale, constituaient une solution possible à ce problème.

L'étude a porté sur 11 régions économiques réparties à travers tout le Canada. On a créé les strates en utilisant des données du recensement de 1971, et on les a évaluées en utilisant les données du recensement de 1981. Pour faire la stratification, on a choisi comme unité de stratification les secteurs de dénombrement du recensement de 1971, sauf au Québec et en Ontario. Pour ces deux provinces on a choisi les sous-divisions de recensement, car le grand nombre de S.D. dans certaines R.É. (jusqu'à 400) aurait occasionné des coûts d'exécution des programmes informatiques trop élevés.

On a utilisé une filière de conversion entre les unités géographiques des deux recensements, pour faire l'évaluation basée sur le recensement de 1981. Les indices basés sur les données de 1981 ont été jugés plus au point pour l'évaluation, étant donné qu'en réalité l'âge moyen des données de stratification sera de 7 ou 8 ans pendant la vie du plan de sondage. On retrouve dans le tableau 2 les indices basés sur les deux recensements.

Tableau 1:

Options de Stratification selon les Variables

	Option de stratification				
Variables	1	2	3	4	5
industries (7)[1]	x	x	x	x	x
revenu		x	x	x	x
employés		x	x		x
chômeurs		x	x*		x
démographie (2)[2]				x	x
logement (4)[3]				x	x
éducation (1)[4]				x	x

[1] nombre de personnes employées dans les secteurs agriculture, forêt et pêche, mines, manufacture, construction, transport, services

[2] population 15-24 ans , population 55 ans et plus

[3] ménages à 1 personne, ménages à 2 personnes, logements possédés, loyer brut total

[4] personnes ayant une éducation secondaire

* poids double pour nombre de chômeurs.

Pour cette étude on a choisi de former des strates contiguës et compactes, en utilisant les vecteurs de contiguité et les variables centroïdes avec des poids moyens de trois (voir sous-section 3.2.2). Le nombre de strates à former par R.É. était le même pour toutes les options.

Tableau 2: Indices de Stratification pour les 5 Options

Variables de Stratification	chômeurs				employés				revenu			
	global		rural/urbain		global		rural/urbain		global		rural/urbain	
	1971	1981	1971	1981	1971	1981	1971	1981	1971	1981	1971	1981
1. 7 industries	5.4	0.1	9.9	3.8	2.9	0.5	8.9	4.8	7.4	5.7	18.9	9.5
2. 7 industries + rev. + emp. + chom.	5.2	2.3	10.2	3.4	8.8	2.7	8.6	3.2	11.2	6.8	22.1	5.9
3. 7 industries + rev. + emp. + chom. × 2	7.4	2.3	10.2	5.3	9.1	2.8	13.1	2.2	10.3	6.8	28.3	9.5
4. 17 variables	6.3	6.4	11.3	4.7	14.1	7.8	12.2	6.4	10.5	9.4	24.4	11.9
5. 15 variables (sauf emp. + chom.)	3.6	0.1	9.8	9.0	6.3	1.6	11.4	3.7	21.0	5.3	28.9	4.5

Tableau 2: Indices de Stratification pour les Options (suite)

Variables de Stratification	agriculture				manufacture			
	global		rural/urbain		global		rural/urbain	
	1971	1981	1971	1981	1971	1981	1971	1981
1. 7 industries	7.4	9.7	37.0	26.0	14.7	8.5	16.9	13.2
2. 7 industries + rev. + emp. + chom.	7.6	7.8	40.0	28.7	10.9	6.6	16.5	12.1
3. 7 industries + rev. + emp. + chom. × 2	8.6	7.9	43.2	31.0	5.5	4.3	14.8	16.1
4. 17 variables	6.1	1.1	40.3	31.8	12.5	13.5	13.3	10.7
5. 15 variables (sauf emp. + chom.)	7.0	0.4	42.7	29.0	7.2	1.4	14.1	16.4

On a tiré les conclusions suivantes des résultats de l'étude qui sont résumés dans le tableau 2:

Type de Stratification: La stratification rural/urbain était de beaucoup supérieure à la stratification globale dans le cas de la variable "agriculture", ce qui n'est pas étonnant. Le même phénomène s'est produit pour la variable "manufacture", quoique moins spectaculaire. Pour la variable "revenu" la stratification rural/urbain était également meilleure, au départ mais elle n'était pas très robuste (i.e., l'indice se détériore avec le temps). La stratification rural/urbain était meilleure pour la variable "chômeurs", tandis qu'il n'y avait pas beaucoup de différence pour "employés".

Variables de stratification: L'option 4, en combinaison avec la stratification rural/urbain était nettement supérieure pour la variable "chômeurs". En ce qui concerne les autres variables, l'option 5 était légèrement meilleure que les autres pour "employés" et "revenu".

3.2.2 Étude sur la contiguïté

Comme mentionné auparavant, on a décidé de retenir l'idée des strates contiguës pour l'E.P.A. Des telles strates devraient être meilleures pour la production d'estimations pour les petites régions, parce que l'échantillon sera bien dispersé au point de vue géographique. De plus on a pensé que les strates contiguës préserveraient davantage l'efficacité du plan d'échantillonnage sur une longue période de temps.

La question a été ensuite de savoir comment utiliser les centroïdes ou les vecteurs de contiguïté, ou les deux ensemble, pour obtenir des strates compactes et contiguës sans que les contraintes géographiques ne dominent trop la minimisation des autres variables.

L'étude a été réalisée avec les mêmes 11 régions économiques. Comme prévu, l'utilisation des seuls vecteurs de contiguïté a résulté en des strates contiguës, mais montrant des configurations souvent irrégulières. D'un autre côté,

l'utilisation des centroïdes seulement, même si des poids très élevés leur ont été associés, n'a donné aucune garantie de contiguïté absolue. En faisant varier les poids des centroïdes, comparativement aux autres variables, on a trouvé que l'utilisation des centroïdes avec poids égal à 3 et des vecteurs de contiguïté ont fourni un bon compromis entre la compacité et l'optimisation non-géographique.

3.3 Nouvelle conception: Stratification

A la lueur des résultats décrits précédemment et aussi en tenant compte des résultats supérieurs d'un plan d'échantillonnage utilisant une stratification rural/urbain tirés d'une étude sur les variances et coûts (Choudhry, Lee, Drew 1985), on a décidé que dans la mesure du possible, on devait procéder à une stratification séparée pour toutes les régions économiques. Des restrictions s'appliquent cependant lorsque dans une R.É. la population rurale, d'une part, ou la population urbaine, d'autre part, n'est pas suffisante pour former au moins une strate. On a déterminé qu'une strate devait pouvoir donner un échantillon d'au moins 90 logements, correspondant à la sélection de deux U.P.É. d'un rendement minimal de 45 logements chacune. Dans le cas où cette contrainte ne pouvait être respectée, on a choisi de procéder à une stratification globale et donc de former des strates mixtes formées de S.D. ruraux et urbains. Ce critère a mené à l'adoption des strates séparées dans plus de 2/3 des R.É.

En ce qui concerne les variables de stratification, on a choisi comme compromis une stratification basée sur les 15 variables de l'option 4 plus "employés". On a ajouté "employés" parce que son inclusion dans l'option 4, comparativement à l'option 5, a amélioré la performance des deux caractéristiques "employés" et "revenu". Suivant la même logique on a exclu "chômeurs" comme variable de stratification.

Pour les contraintes géographiques, on a décidé d'utiliser les vecteurs de contiguïté en combinaison avec un poids uniforme de centroïdes égal à 3 dans toutes les régions économiques.

On devait également prendre une décision à propos du nombre de strates par R.É.. En pratique, dans la plupart des cas on n'a pas eu le choix. Selon le plan d'échantillonnage, chaque U.P.É. correspond à une tâche d'interviewer, et on désirait sélectionner au moins deux U.P.É. par strate afin de permettre l'estimation de la variance sans biais. Étant donné ces contraintes, dans près de 2/3 des cas on n'a formé qu'une seule strate avec 2 ou 3 U.P.É. sélectionnées, dans les parties urbaines, rurales ou une combinaison des deux. Dans les autres cas, on a stratifié de façon à sélectionner également 2 ou 3 U.P.É. par strate. On s'est basé sur une autre étude montrant de légères réductions dans la variance pour cette façon de procéder, comparativement à l'ancien plan de sondage où on sélectionnait de 4 à 6 U.P.É. par strate (Choudhry, Lee, et Drew 1985).

3.4 Étude sur la robustesse des strates contiguës et non-contiguës

Des strates robustes sont des strates qui maintiennent l'efficacité du plan d'échantillonnage au fil du temps. Après le remaniement on a entrepris une étude pour voir si les strates contiguës seraient plus robustes, comme on en a fait l'hypothèse.

L'étude a porté sur trois régions économiques de l'Ontario, les R.É. 520, 540 et 580 (selon la numérotation de 1981). Pour chacune de ces régions, on a comparé les résultats de la nouvelle stratification (choisie pour le remaniement de l'E.P.A.) qui consiste en des strates contiguës, avec une stratification sans contrainte de contiguïté. On a stratifié en utilisant les données de 1981, et on a fait l'évaluation avec les données de 1971. Pour les strates contiguës on a utilisé les vecteurs de contiguïté avec centroïdes, tandis que pour les strates non-contiguës on a essayé deux options donnant des poids de 0 et 3 aux centroïdes, respectivement. Les variables de stratification utilisées ont été les mêmes 16 variables que l'on a déjà décrites (soit l'option 4 modifiée).

Les résultats se retrouvent dans le tableau 3. On voit qu'en général l'indice global calculé lors de la stratification est plus élevé pour les deux options où la contiguité n'est pas nécessaire, comme on pouvait s'y attendre (colonne

1981). Cependant, ces deux options donnent également des indices plus élevés au fil du temps (colonne 1971).

Tableau 3

Indices de Stratification selon les contraintes géographiques

Région Économique	No. de Strates	Contraintes Géographiques					
		Contiguïte et Centroïdes (poids de 3)		Centroïdes (poids de 3)		Aucune	
		1981	1971	1981	1971	1981	1971
520	2	32.2	28.5	30.2	30.1	34.5	27.0
540	3	21.8	14.1	24.9	17.8	35.2	26.8
580	4	22.8	18.9	41.4	33.7	43.7	38.5

A-t-on vraiment besoin de strates contigües? Pour répondre à cette question, on devrait entreprendre une étude plus approfondie comportant des R.É. de plusieurs provinces. L'évaluation de la robustesse de la stratification poserait alors certains problèmes. Il est facile d'évaluer la robustesse en Ontario car la stratification y est faite au niveau des sous-divisions de recensement dont les frontières n'ont à peu près pas changé depuis 1971. A l'opposé, lorsque la stratification est faite au niveau des secteurs de dénombrement, qui ont des frontières changeantes d'un recensement à l'autre, il est très difficile d'obtenir des chiffres comparables entre recensements en ce qui concerne la robustesse, surtout lorsque les strates ne sont ni compactes ni contiguës.

S'il devait s'avérer que la stratification sans contiguïté est plus optimale, cela pourrait compenser pour les problèmes éventuels rencontrés dans l'estimation pour les petites régions. Cela pourrait également ouvrir de nouveaux horizons: en effet, une fois libérés des contraintes de contiguïté, pourquoi ne pas former tout d'abord les U.P.É., compactes mais pas nécessairement contiguës, pour les regrouper seulement ensuite en des strates? Encore une fois, on ne peut répondre qu'en procédant à de nouvelles études plus approfondies.

3.5 Formation des U.P.É.

On a modifié l'algorithme de classification pour effectuer la formation des U.P.É. dans les strates rurales et mixtes. Dans les strates rurales en particulier, la formation des U.P.É. ressemble beaucoup à la stratification, au point de vue conception. La seule différence tient au fait que dans la stratification, on veut minimiser la somme des carrés des variables géographiques et non-géographiques à l'intérieur de chaque strate, tandis que dans la formation des U.P.É. on veut minimiser la somme des carrés des variables géographiques (de façon à obtenir des U.P.É. compactes pour réduire les coûts) et maximiser celle des variables non-géographiques. Ce dernier critère permet d'obtenir les U.P.É. les plus hétérogènes possible en ce qui a trait aux caractéristiques, de façon à ce qu'elles soient toutes bien représentatives de la strate au moment de l'échantillonnage.

Il y a cependant un conflit entre la compacité désirée des U.P.É. et leur hétérogénéité, à cause de la tendance des unités adjacentes à posséder des caractéristiques similaires. Étant donné les coûts d'ordinateur peu élevés, on a fait 3 délimitations par strate avec des poids sur les centroïdes de 10, 15 et 20, relativement aux autres variables. On a ensuite montré les résultats de chaque délimitation sur un graphique dont les axes représentent les centroïdes (voir figure 1). On a fait un choix entre les 3 délimitations en tenant compte de la qualité de l'optimisation des variables, reflétée par les indices de stratification, et en consultant les graphiques. On tenait également compte d'un indice de compacité. En pratique, on a choisi la plupart du temps un poids de 10 ou de 15 sur les centroïdes.

La formation des U.P.É. dans les strates mixtes a amené une contrainte supplémentaire. On désirait en effet que la proportion de la population urbaine soit à peu près la même dans chaque U.P.É. Étant donné qu'on voulait de plus avoir des U.P.É. de populations totales approximativement égales, il a donc été nécessaire parfois de partager les grands centres urbains parmi plus d'une U.P.É. La solution retenue a été la suivante:

FIGURE 1

Exemple de délimitation d'U.P.É.

Chaque unité de stratification est représentée par une lettre qui identifie l'U.P.É. à laquelle appartient l'unité. On a encerclé les U.P.É. pour mieux les différencier.

1. On détermine le nombre de parties de centres urbains que recevra en moyenne une U.P.É. (N). Ce nombre dépend de la proportion de la population urbaine dans la strate ainsi que du nombre d'unités urbaines. En pratique, il a été fixé à 1 ou 2. Certaines strates ne présentant pas une population ou un nombre d'unités urbaines suffisantes se sont vues reclassifiées comme étant des strates entièrement rurales.

2. On détermine le nombre de parties par lesquelles chaque centre urbain sera divisé. Le nombre total de parties doit égaler N fois le nombre d'U.P.É. et chaque centre urbain est divisé en un nombre de parties proportionnel à sa population.

3. On applique le programme de stratification optimale en considérant chaque partie de centre urbain comme une unité de stratification distincte et en ajoutant la variable "population urbaine" aux autres variables de stratification. On ajuste le poids accordé à cette variable pour obtenir une répartition rural/urbain la plus égale possible dans chaque U.P.É., en essayant de ne pas trop affecter la compacité et l'optimisation globale. Ceci est réalisé par tâtonnement seulement. On a trouvé en pratique qu'un poids de 10 ou 15 sur la population urbaine, relativement aux autres variables, menait à des résultats satisfaisants.

Dans les strates urbaines, les U.P.É. ont été formées de centres urbains. On a parfois combiné des centres petits et peu éloignés, mais sans considération d'optimalité des caractéristiques.

Le tableau 4 donne les indices de délimitation moyens pour les U.P.É. dans les strates rurales, mixtes et urbaines. Pour les variables non-géographiques, l'indice le plus bas représente la meilleure délimitation, tandis que c'est le contraire pour les centroïdes. Les résultats sont évidement meilleurs, au point de vue optimalité des caractéristiques, pour les strates rurales et mixtes, où on a utilisé l'algorithme de classification. Les indices élevés des centroïdes révèlent que les U.P.É. sont bien compactes.

Tableau 4

Indices Moyens de Délimitation des U.P.É.

Variables	Type de strate		
	rurale	mixte	urbaine
agriculture	8.1	8.3	9.0
foresterie	21.8	24.5	35.9
mines	20.6	36.0	57.0
manufacture	15.1	22.9	53.3
construction	9.0	11.4	22.7
transport	9.9	12.8	22.7
services	9.4	12.8	29.1
employés	7.7	10.2	23.6
chômeurs*	13.6	14.2	18.6
revenu	8.9	11.2	23.7
population 15-24	9.4	13.4	29.8
population 55+	7.4	13.9	34.5
ménages 1 personne	5.1	7.4	13.0
ménages 2 personnes	7.9	11.9	28.1
logements possédés	6.8	12.5	29.4
loyer brut total	5.1	7.7	14.4
éducation secondaire	9.1	10.5	17.4
population totale*	3.2	4.0	10.5
logements*	5.9	8.9	18.6
centroïde 1	91.6	92.7	99.2
centroïde 2	90.5	91.7	97.2

* non utilisé comme variable dans l'optimisation.

4. STRATIFICATION DANS LES UNITÉS AUTOREPRÉSENTATIVES

4.1 Ancienne conception

Les unités autoreprésentatives de l'ancien plan de sondage correspondaient aux villes qui étaient assez grandes pour avoir comme rendement prévu un échantillon suffisant pour être traité par un interviewer. La limite inférieure de la taille des U.A.R. variait de 10,000 personnes dans les provinces de l'Atlantique à 29,000 personnes au Québec et en Ontario.

Les grandes U.A.R. étaient stratifiées géographiquement en regroupant de 3 à 5 secteurs de recensement (S.R.) contigus, sans recherche d'optimalité. Les S.R.sont des unités géostatistiques comprenant de 3,000 à 5,000 habitants, et dont la stabilité d'un recensement à l'autre en fait des unités opérationnelles pratiques. On s'attendait à ce que ces strates soient efficaces pour l'estimation des caractéristiques, et que leur petite taille (entre 10,000 et 15,000 personnes) permette la mise à jour de l'échantillon dans les secteurs expérimentant une croissance rapide, sans d'autre part déranger l'échantillon.

En plus de la base de sondage aréolaire, une base ouverte était tenue à jour pour les immeubles d'appartements dans les grandes villes.

4.2 Étude sur la stratification

On a considéré trois grandes U.A.R. dans cette étude, soient Québec, Ottawa et Toronto. L'unité de stratification choisie a été le secteur de recensement. A cause de contraintes opérationnelles imposées par le programme de stratification, on a dû séparer Toronto en six parties correspondant généralement aux grandes divisions naturelles de la ville. Toute stratification a été menée séparément dans chacune de ces parties. On a utilisé les mêmes 16 variables de stratification qui ont finalement été choisies dans la partie N.A.R.

Deux options principales ont été évaluées:

Option 1: Stratification à deux niveaux:

- strates primaires contiguës et compactes, avec un poids de 3 sur les centroïdes et une taille déterminée en vue d'un rendement d'environ 150 logements.

- strates secondaires - 4 ou 5 par strates primaires, formulées sans contraintes géographiques.

Option 2: stratification compacte formulée avec l'emploi de centroïdes (poids de 3) sans utiliser les vecteurs de contiguïté, avec une taille comparable à celle des strates secondaires de l'option 1.

Le tableau 5 montre les résultats de la comparaison entre l'ancienne stratification et les deux options étudiées. Comme dans la partie N.A.R., la stratification a été effectuée avec les données du recensement de 1971, pour ensuite être évaluée avec celles du recensement de 1981.

On voit que les deux options étudiées donnent toujours de meilleurs indices que l'ancienne stratification, sauf peut-être pour les trois premières variables dont l'importance est de toute façon faible dans les villes. L'ancienne stratification est quand même très valable si l'on considère qu'elle a été faite sans souci d'optimalité.

On remarque également que les trois méthodes donnent une stratification généralement robuste au fil du temps, telle que réflétée par la comparaison entre les indices 1981 et 1971. Des exceptions importantes à cette règle, malheureusement, semblent être les caractéristiques employés et chômeurs.

Tableau 5

Comparaison de trois méthodes de stratification (U.A.R.)

	Ancienne conception		Stratification à deux niveaux (Option 1)		Stratification Compacte (Option 2)	
Variables	1971	1981	1971	1981	1971	1981
agriculture	5.5	2.9	3.2	.8	3.4	1.8
foresterie	2.2	2.3	2.1	1.7	2.2	2.3
mines	7.6	4.9	8.5	4.1	7.6	4.0
manufacture	34.7	35.0	36.6	34.1	39.1	35.0
construction	32.5	29.6	39.7	30.1	42.4	33.4
transport	9.2	6.8	18.0	11.6	20.0	11.6
services	29.5	27.5	45.8	33.1	46.7	32.1
employés	15.1	8.0	31.4	14.1	32.8	12.6
chômeurs*	14.6	5.7	14.9	6.7	15.5	7.1
revenu	39.4	38.6	51.8	29.8	53.6	48.0
population 15-24	9.6	15.2	12.5	17.5	13.3	14.9
population 55+	27.9	18.3	34.0	20.8	32.6	18.5
ménages 1 personne	20.3	19.2	36.3	33.8	37.8	35.0
ménages 2 personnes	21.9	20.3	40.3	30.9	40.1	30.2
logements possédés	20.3	15.3	29.7	22.9	32.1	24.9
éducation secondaire	32.6	42.4	50.3	47.9	51.6	49.1
population 15+*	27.0	8.2	38.0	13.4	37.6	12.0
logements*	21.8	18.5	41.7	33.8	42.1	34.3

* non utilisé comme variable de stratification.

4.3 Nouvelle conception

Étant donnée la similitude des résultats entre les deux options étudiées, on a décidé d'adopter la stratification à deux niveaux (option 1) dans les grandes villes où l'échantillon est de 300 ménages ou plus, pour les raisons suivantes:

(i) le fait d'avoir la contiguïté au niveau des strates primaires nous donne une unité convenable pour la mise à jour de l'échantillon

(ii) les strates primaires pourront être utilisées pour la formation des tâches d'interviewer. La taille des strates a été déterminée de manière à ce que l'échantillon à l'intérieur du secteur, soit l'échantillon aréolaire plus l'échantillon provenant des strates d'appartements, corresponde à deux tâches d'interviewer (160 ménages dans le coeur de la ville et 120 ailleurs).

(iii) la stratification à deux niveaux mène à une meilleure représentation de la variance de réponse corrélative dans les estimations de la variance. Dans l'ancien plan de sondage, il n'y avait habituellement qu'un seul interviewer par strate, ce qui occasionnait une sous-estimation de cette composante de la variance. Avec des strates secondaires non-géographiques, mais des tâches d'interviewers toujours géographiques, ce problème sera moins fréquent.

Les contraintes de coûts associées à l'exécution des programmes informatiques impliqués nous ont obligé à traiter certaines U.A.R. de façon particulière. En effet, la région de Montréal se voit diviser en sept parties indépendantes, le temps de la stratification. Il en est de même avec Toronto (5 parties, Winnipeg (2 parties), Calgary (2 parties), Edmonton (2 parties) et Vancouver (3 parties). Ces divisions se sont effectuées à partir de critères "naturels" tels que suggérés par la géographie de ces régions.

Dans les grandes U.A.R., les immeubles d'appartements existant au moment de la conception du plan de sondage ont été triés en fonction des strates primaires

dans lesquelles ils étaient situés, pour atteindre une stratification implicite de cet échantillon.

Dans les U.A.R. de taille moyenne, où l'échantillon était insuffisant pour justifier la stratification à deux niveaux, on s'est contenté de construire des strates optimales à l'aide de l'algorithme de stratification, sans aucune contrainte géographique.

Les plus petites U.A.R., celle qui n'ont pas été décomposées en côtés d'îlot aux fins du recensement, ont été stratifiées manuellement, sans souci d'optimalité.

Mentionnons finalement que la période d'introduction du nouvel échantillon nous a amené une contrainte supplémentaire. Pour les grandes U.A.R., on a défini le secteur noyau comme étant constitué des strates complètes de l'ancienne conception qui n'ont pas été affectées par des changements de frontières. En s'assurant que les frontières des strates de la nouvelle conception respectent ces secteurs noyau, on s'est aussi assuré que le nouvel échantillon représente la même aire géographique que l'ancien échantillon durant la période d'introduction. Ceci a permis le remplacement graduel de l'ancien échantillon par le nouveau et par le fait même évité un coûteux chevauchement du nouvel échantillon et de l'ancien échantillon (Mayda, Drew, et Lindeyer 1985).

5. CONCLUSION

L'utilisation de l'algorithme de classification multi-variée nous a permis de développer une stratification très générale, renforçant ainsi l'E.P.A. dans son rôle d'enquête-ménages générale. De plus, l'automatisation des étapes de stratification dans les parties N.A.R. et A.R., et de la délimitation des U.P.É. dans les U.N.A.R., a mené à une réduction significative dans le coût et le temps nécessaire pour remanier l'échantillon.

Le système est documenté (Foy 1984) et il peut être utilisé pour la formation

de strates ou de grappes dans les autres enquêtes. On pourrait également s'en servir dans des situations où on a à définir des régions statistiques ou administratives, en utilisant toute une gamme de variables.

Pour l'E.P.A., un aspect de recherche à approfondir serait le choix entre des strates contiguës et non-contiguës, et les implications des strates non-contiguës sur le plan d'échantillonnage.

REMERCIEMENTS

Les auteurs désirent remercier Sylvie Trudel et Marc Joncas pour leur aide apportée à la réalisation des études mentionnées dans ce rapport, ainsi que les membres du Comité du remaniement de l'E.P.A. pour leurs précieuses suggestions. Ils remercient aussi l'arbitre pour ses commentaires utiles.

BIBLIOGRAPHIE

CHOUDHRY, G.H., LEE, H., et DREW, J.D. (1985). Cost-variance optimization for the Canadian labour force survey. American Statistical Association Proceedings of the Section on Survey Research Methods.

DAHMSTRÖM, P., et HAGNELL, M. (1978) The formation of strata using cluster analysis. Document interne, Department of Statistics, University of Lund, Sweden.

FOY, P. (1984). Programme de stratification pour l'enquête sur la population active du Canada: Guide de l'usager. Document interne, division des méthodes de recensement et d'enquêtes ménages, Statistique Canada.

FRIEDMAN, H.P., et RUBIN, J. (1967). On some invariant criteria for grouping data. Journal of the American Statistical Association, 62, 1159-1178.

JUDKINS, D.R., et SINGH, R.P. (1981). Using clustering algorithms to stratify primary sampling units. American Statistical Association Proceedings of the Section on Survey Research Methods, 274-284.

KOSTANICH, D., JUDKINS, D.R., SINGH, R.P., et SCHANTZ, M. (1981). Modification of Friedman-Rubin's clustering algorithm for use in stratified PPS sampling. American Statistical Association Proceedings of the Section on Survey Research Methods,

MAYDA, F., DREW, J.D., et LINDEYER, J. (1985). Phase-in of the redesigned Labour Force Survey. Document interne, division des méthodes de recensement et d'enquêtes ménages, Statistique Canada.

PLATEK, R., et SINGH, M.P. (1976). Méthodologie de l'enquête sur la population active du Canada. No. 71 - 526 au catalogue, Statistique Canada.

SINGH, M.P., DREW, J.D., et CHOUDHRY, G.H. (1984). Remaniement de l'enquête sur la population active au Canada à partir des résultats du recensement de 1981. Technique d'enquête, 10, 139-154.